STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2407

Multiple hypothesis testing strategies for genetic case–control association studies [¶]

Philip S. Rosenberg[*,†], Anney Che[‡] and Bingshu E. Chen[§]

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, MD, U.S.A.

SUMMARY

The genetic case–control association study of unrelated subjects is a leading method to identify single nucleotide polymorphisms (SNPs) and SNP haplotypes that modulate the risk of complex diseases. Association studies often genotype several SNPs in a number of candidate genes; we propose a two-stage approach to address the inherent statistical multiple comparisons problem. In the first stage, each gene's association with disease is summarized by a single p-value that controls a familywise error rate. In the second stage, summary p-values are adjusted for multiplicity using a false discovery rate (FDR) controlling procedure. For the first stage, we consider marginal and joint tests of SNPs and haplotypes within genes, and we construct an omnibus test that combines SNP and haplotype analysis. Simulation studies show that when disease susceptibility is conferred by a SNP, and all common SNPs in a gene are genotyped, marginal analysis of SNPs using the Simes test has similar or higher power than marginal or joint haplotype analysis. Conversely, haplotype analysis can be more powerful when disease susceptibility is conferred by a haplotype. The omnibus test tracks the more powerful of the two approaches, which is generally unknown. Multiple testing balances the desire for statistical power against the implicit costs of false positive results, which up to now appear to be common in the literature. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: statistics and numerical data; research design; epidemiology; case–control studies; human genome; polymorphism; single nucleotide; haplotypes

1. INTRODUCTION

Following the completion of the human genome reference sequence, large-scale resequencing projects have identified large numbers of polymorphisms [1] in the form of single nucleotide

[*] Correspondence to: Philip S. Rosenberg, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Executive Plaza South, Room 8022, Rockville, MD 20852-7244, U.S.A.

[†] E-mail: rosenbep@mail.nih.gov

[‡] E-mail: chea@mail.nih.gov

[§] E-mail: cheneric@mail.nih.gov

[¶] This article is a U.S. Government work and is in the public domain in the U.S.A.

Contract/grant sponsor: NIH

Received 14 October 2004
Accepted 1 September 2005

polymorphisms (SNPs) [2] and SNP haplotypes [3]. In principle, this knowledge allows studies to associate genetic polymorphisms with disease informed by a comprehensive understanding of human genetic variation [4–9]. Whole-genome association studies have also been proposed [10–12].

Despite the possibilities of whole-genome analysis, for reasons of cost and statistical efficiency [13], case–control analysis of specific candidate genes will likely remain a mainstay approach. It is clear, however, that many studies will consider, if not the whole genome, at least whole panels of genes, for example, panels of DNA repair genes or genes that function in specific aetiological pathways.

As SNP association data accumulates, statisticians will increasingly be called upon to assist in the analysis. At our institution, a number of genetic case–control association studies are in progress. The data sets of these studies typically contain at least a few SNPs in each of several candidate genes. As genotyping technologies advance, it may become cost-effective to genotype all common SNPs in candidate gene or gene region. The SNPs within each gene are typically correlated because of linkage disequilibrium (LD). Often, the correlation is high, making it difficult to discern which of a number of tightly linked SNPs might be directly associated with case–control status. When this is the case, one logical focus of a SNP analysis is to detect whether any SNP is associated with case–control status. The SNP variants in different genes may or may not be correlated, depending on the relative positions of the genes in the genome.

This organization—numerous SNPs within genes, and genes within panels—naturally lends itself to a two-stage analysis. In the first-stage analysis of each candidate gene, it is desirable to summarize the evidence for association using a summary *p*-value that combines evidence for association over a number of variants. For this purpose, there is interest both in SNP-based associations and haplotype-based associations [14–18]; frequently, both approaches are investigated. Ideally, the summary *p*-value should reflect both analytical approaches if both are performed. In the second-stage analysis of gene panels, it may be desirable to adjust the summary gene-level *p*-values for multiple comparisons, using an approach that controls an appropriate type I error rate. In both stages, there is an inherent multiple comparisons problem, one that has been widely recognized to be a looming issue [19–21].

In this study, we adapt some modern multiple comparisons procedures [22] to tackle the multiplicity issue inherent in each stage of the analysis. For the first stage, we adapt the Simes test [23] to the problem of combining multiple single-locus marginal SNP tests within a candidate gene, and we construct an omnibus test that combines the Simes test of SNPs with a joint test of haplotypes. For the second stage, we consider false discovery rate (FDR) controlling multiple comparisons procedures [24]. To explore the operating characteristics of these tests, we develop an approach to simulate case–control studies of a candidate gene that incorporate empirical haplotype frequencies and patterns of LD, and we evaluate the power of testing procedures using simulation studies conducted over a panel of nine candidate genes.

2. METHODS

2.1. *Multiple testing strategies for a single candidate gene*

A primary goal of many candidate gene studies is to identify whether *any* sequence variation in a queried gene or gene region is associated with disease. To avoid obtaining a false negative result, many studies consider both SNP- and haplotype-based associations. SNP-based associations are biologically plausible on the causal hypothesis and on the proximity hypothesis, i.e. because the tested SNP has a direct effect on disease susceptibility or because it is in LD with another SNP, insertion or deletion polymorphism, etc. that does [25]. Haplotype-based associations may also be biologically plausible, because haplotypes encode the units of transcription. The two approaches are not equivalent because SNPs frequently travel on more than one common haplotype. Whichever approaches are considered, it appears desirable to control the familywise error rate (FWER), defined as the probability of falsely declaring that any tested feature in the gene is associated with case–control status.

Therefore, we first consider four testing approaches: marginal and joint tests of SNPs, and marginal and joint tests of haplotypes. For haplotype analyses, our primary simulation studies will evaluate ideal tests that incorporate phase-known haplotype data, because this defines the best-case scenario for haplotype analyses. In selected situations, we study haplotype analysis of unphased SNP genotype data, using a method described by Lake *et al.* [26] to account for phase ambiguity.

To define the statistical models and test statistics, let $Y_i, i=1,\ldots,n_1,n_1+1,\ldots,n_1+n_0$ equal 1 for the n_1 cases and 0 for the n_0 controls, so that $n=n_1+n_0$ is the total sample size. Consider a gene with m SNPs and p haplotypes. Let $\mathbf{X}^{\mathrm{G}}$ be the genotype scoring matrix with rows $\mathbf{x}_i^{\mathrm{G}}, i=1,\ldots,n$ and elements $x_{i,j}^{\mathrm{G}}, i=1,\ldots,n, j=1,\ldots,m$ such that $x_{i,j}^{\mathrm{G}}=0,1$, or 2 if subject i has $0,1$, or 2 copies of the variant allele for SNP j. It is arbitrary which allele is scored as the variant. Let $\mathbf{X}^{\mathrm{H}}$ be the haplotype scoring matrix with rows $\mathbf{x}_i^{\mathrm{H}}$, $i=1,\ldots,n$ and elements $x_{i,k}^{\mathrm{H}}$, $i=1,\ldots,n, k=1,\ldots,p$ such that $x_{i,k}^{\mathrm{H}}=0,1$, or 2 if subject i has $0,1$, or 2 copies of haplotype k. Because each subject is diploid, $\mathbf{X}^{\mathrm{H}}\mathbf{1}_p=2\mathbf{1}_n$, where $\mathbf{1}_l$ denotes a column vector of l ones.

For SNP-based analysis, we consider a joint co-dominant logistic model for SNPs

$$M_{\mathrm{Joint}}^{\mathrm{SNP}}: \operatorname{logit} P(Y_i=1|\mathbf{x}_i^{\mathrm{G}}) = \alpha_0 + \beta_1 x_{i,1}^{\mathrm{G}} + \cdots + \beta_m x_{i,m}^{\mathrm{G}}$$

and test for the significance of the m SNPs *versus* the null model with only an intercept term by comparing the corresponding likelihood ratio test statistic against a χ^2 distribution with m degrees of freedom. More generally, the model might adjust for design and environmental factors. The β-coefficients measure the change in the log odds of disease per copy of each SNP, controlling for all other SNPs. We call the p-value obtained from this procedure $p_{\mathrm{Joint}}^{\mathrm{SNP}}$.

If a single SNP is causally associated with disease, the joint model and its omnibus test on m degrees of freedom may not be very powerful. For these scenarios, a marginal test may be more sensitive. Therefore, we consider the corresponding sequence of marginal models

$$\{M_j^{\mathrm{SNP}}\}_{j=1}^m: \operatorname{logit} P(Y_i=1|x_{i,j}^{\mathrm{G}}) = a_{0,j} + b_j x_{i,j}^{\mathrm{G}}$$

and the corresponding sequence of single-SNP likelihood ratio tests, each on one degree of freedom. The b-coefficients measure the change in the log odds of disease per copy of each SNP, marginally over all other SNPs. For each SNP we obtain a two-sided p-value. From the complete array of m SNP genotypes, we obtain a realization of the ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$, with p_1 the p-value for the most significant SNP and $p_{(m)}$ the p-value for the least significant SNP. To control the gene-wide type I error rate, we apply the Simes test [23] to the array of single-SNP trend tests. Formally, define

$$p_{\mathrm{Marg}}^{\mathrm{SNP}} = \min\left(1, \min_{1\leqslant k}\left\{p_{(k)}\frac{m}{k}\right\}\right)$$

this expression equals the Simes adjusted p-value for the most significant SNP. At level α, the Simes test rejects the complete null hypothesis if $p_{\mathrm{Marg}}^{\mathrm{SNP}} \leqslant \alpha$. The Simes test can be more powerful than the classical Bonferroni test, especially when the test statistics are positively correlated [23]. When there is both positive and negative correlation, the Simes test may be too liberal, but studies suggest that the exceedance of the nominal level will usually be modest [27–29].

The logic underlying the haplotype-based tests is similar. We consider a joint co-dominant logistic model for haplotypes

$$M_{\mathrm{Joint}}^{\mathrm{Hap}}: \operatorname{logit} P(Y_i = 1|\mathbf{x}_i^{\mathrm{H}}) = \gamma_0 + \delta_2 x_{i,2}^{\mathrm{H}} + \cdots + \delta_p x_{i,p}^{\mathrm{H}}$$

and test for its significance by comparing the corresponding likelihood ratio test statistic against a χ^2 distribution with $p-1$ degrees of freedom. The haplotype scored in the first column of $\mathbf{X}^{\mathrm{H}}$ serves as the referent haplotype. The δ-coefficients measure the change in the log odds of disease for subjects with one copy of the corresponding haplotype and one copy of the referent haplotype, compared to persons with two copies of the referent haplotype. The observed p-value associated with this procedure is $p_{\mathrm{Joint}}^{\mathrm{Hap}}$.

Similarly, we consider the corresponding sequence of marginal models

$$\{M_k^{\mathrm{Hap}}\}_{k=1}^{p}: \operatorname{logit} P(Y_i = 1|x_{i,k}^{\mathrm{H}}) = c_{0,k} + d_k x_{i,k}^{\mathrm{H}}$$

and the corresponding sequence of single-haplotype likelihood ratio tests, each on one degree of freedom. The d-coefficients measure the change in log odds of disease for persons with one copy of the corresponding haplotype and one copy of any other haplotype, compared to persons with two copies of any other haplotype. We adjust the p single-haplotype tests for multiplicity using the Simes testing approach, obtaining $p_{\mathrm{Marg}}^{\mathrm{Hap}}$, the summary adjusted p-value for the most significant haplotype.

In practice, the genetic mechanism conferring disease susceptibility may be uncertain. For this situation, we propose a composite omnibus test using the test statistic

$$\mathrm{omni} = \min(p_{\mathrm{Marg}}^{\mathrm{SNP}}, p_{\mathrm{Joint}}^{\mathrm{Hap}})$$

In this test, we combine a marginal model for SNPs with a joint model for haplotypes; these models are often of specific epidemiological interest *a priori*. In general, other models could be combined, although use of the joint model for SNPs requires that subjects have complete genotype data for all SNPs. We compute the distribution of the omnibus test statistic under the complete null hypothesis from the permutation distribution obtained by shuffling case and control indicators. We reject the null if the observed value of omni is less than or equal to the α-level quantile of the null distribution. For computational efficiency, we also consider an approximate omnibus test obtained by applying the Bonferroni correction to the two p- values $p_{\mathrm{Marg}}^{\mathrm{SNP}}$ and $p_{\mathrm{Joint}}^{\mathrm{Hap}}$. The approximate omnibus test statistic is

$$\tilde{p}_{\mathrm{omni}} = \min(1, \min(2p_{\mathrm{Marg}}^{\mathrm{SNP}}, 2p_{\mathrm{Joint}}^{\mathrm{Hap}}))$$

and we reject the null if $\tilde{p}_{\mathrm{omni}} \leqslant \alpha$.

2.2. Multiple testing strategies for a panel of candidate genes

To test a panel of G candidate genes in a single study, one might choose to apply a FWER procedure to the summary p-values obtained from the first-stage analysis, as described elsewhere [30]. We propose here to control the expected FDR using the Benjamini–Hochberg FDR procedure (BH-FDR) [24].

Specifically, let $p_{\text{Marg}}^{\text{SNP}}, p_{\text{Joint}}^{\text{SNP}}, p_{\text{Marg}}^{\text{Hap}}, p_{\text{Joint}}^{\text{Hap}}$, or p_{omni} be summary gene-wide p-values for candidate genes $g, g=1,\ldots,G$, as defined previously. Generically, let the summary p-value for gene $g, g=1,\ldots,G$ be p^g. To control the expected FDR at the gene level, we apply the BH-FDR procedure to the ordered summary p-values $p^{(1)},\ldots,p^{(G)}$, where $p^{(1)}$ is the p-value of the most significant gene in the panel and $p^{(G)}$ is the p-value of the least significant gene in the panel. The test statistics p^g may or may not be independent under the null; genes with SNPs in LD will not have independent test statistics.

To point to particular genes that are associated with disease, we use the FDR-adjusted p-values [31]

$$\tilde{p}^{(g)} = \min\left(1, \min_{g \leqslant r}\left\{p^{(r)}\frac{G}{r}\right\}\right)$$

and declare a gene to be significantly associated with disease at FDR level q if its adjusted p-value is less than or equal to q. Clearly, the more powerful the FWER procedure we can apply to each gene, the more powerful will be the FDR procedure we can apply to the panel.

The BH-FDR procedure will control the expected FDR for any configuration of genes that are associated with case–control status, so long as the test statistics of the non-associated genes are independent [24] or positively dependent [32, 33]. When analysing genes that are not associated with disease and that are not in LD for any reason (i.e. genes on different chromosome arms or chromosomes), one- or two-sided test statistics should be independent, and the BH-FDR procedure will control the expected FDR.

The situation is more complicated for tightly linked genes that each might have an independent effect. Because SNPs can exhibit both positive and negative LD, one-sided tests will also have positive and negative correlation under the null, and the BH-FDR procedure may or may not provide FDR control. In this situation, the effect of negative LD should be minimized or removed for two-sided tests. However, negatively correlated two-sided tests might be generated if disease susceptibility was modulated by a ‘liability score’ that depended on the number of certain alleles that were present, rather than the specific alleles.

2.3. Haplotype structures and disease incidence models

We consider case–control association studies of unrelated subjects, where individual-level SNP genotype data for all non-redundant SNPs in a candidate gene or gene region are available for analysis. We study the idealized situation in which phase-known SNP haplotype data are also available for analysis. We consider empirical patterns of LD of SNPs within a candidate gene, and the empirical spectrum of haplotype frequencies, obtained from a resequencing study of a panel of nine human candidate genes [34] (Table I).

In this panel, the number of identified SNPs ranges from 6 to 22, the number of non-redundant SNPs with a minor allele frequency $\geqslant 0.05$ ranges from 3 to 10, and the number of haplotypes with a frequency $\geqslant 0.05$ ranges from 3 to 6. For each gene, the haplotype structure can be encoded by a binary matrix $\mathbf{H}$ with rows that correspond to haplotypes and columns

Table I. Characteristics of the nine-gene panel.

Gene	Locus	Number of SNPs	Number of non-redundant SNPs[*]	Number of non-redundant SNPs with minor allele frequency $\geqslant 0.05$[†]	Number of haplotypes	Number of haplotypes with $f \geqslant 0.05$[‡]
CASP8	2q33-q344	13	11	7	12	5
CASP10	2q33-q34	11	7	5	6	4
CFLAR	2q33-q34	6	4	3	4	3
CTLA4	2q33	12	9	9	11	6
GAD2	10q11.23	13	7	5	8	4
H19	11p15.5	13	10	9	15	5
INS	11p15.5	14	8	5	9	4
SDF1	10q11.1	22	11	10	10	6
TCF8	10p11.2	14	13	6	15	6

[*] Redundant SNPs are in perfect LD with other SNPs.

[†] These SNPs are tested in the SNP-based analyses.

[‡] The haplotype frequencies f are here normalized to sum to 1. In the original resequencing studies some haplotypes could not be determined. The normalized frequencies will be less than or equal to the frequencies that were actually observed. For the joint model for haplotypes, rare haplotypes with normalized $f < 0.05$ are pooled.

that correspond to SNPs. One haplotype (usually the most common one) is arbitrarily chosen to be the reference haplotype, and $\mathbf{H}_{h,j} = 1$ if SNP j in haplotype h differs from SNP j in the reference haplotype, 0 otherwise, $h = 1, \ldots, p, j = 1, \ldots, m$. The haplotype frequencies are given by f. Haplotype structures for the nine-gene panel are shown in Figure 1. Of 59 variant SNP alleles with a minor allele frequency $\geqslant 0.05$, 28 (47 per cent) travel on 1 haplotype with a haplotype frequency $\geqslant 0.05$, 14 (24 per cent) travel on 2 such haplotypes, 10 (17 per cent) travel on 3, and 7 (12 per cent) travel on 4 or 5. In terms of $\mathbf{H}$ and f, the proportion of haplotypes carrying variant alleles at position i and j is $\pi_{ij} = \sum_{h=1}^{p} f_h \mathbf{H}_{hi} \mathbf{H}_{hj}$, and the pairwise correlation between SNPs is $D_{ij}^{\mathrm{Corr}} = (\pi_{ij} - \pi_i \pi_j)/[\pi_i(1-\pi_i)\pi_j(1-\pi_j)]^{1/2}$. The genes described in Figure 1 show heterogeneous patterns of positive and negative LD; a number of genes appear to have at least one block of high LD (data not shown).

We assume that disease susceptibility follows a prospective linear logistic model. We consider co-dominant, dominant, and recessive models of SNP and haplotype effects. In simulation studies for each gene, we evaluate scenarios where each SNP and haplotype in turn is associated with disease with a relative risk (RR) of 1.5. The RR for the co-dominant models is increased by the factors 1.5 and $1.5^2 = 2.25$, respectively, in persons who carry one or two copies of the variant allele. Because many pairs of SNPs in these genes are tightly linked, the increased risk conferred by any single SNP induces a complex pattern of association across a number of other SNPs. Similarly, the increased risk conferred by any single haplotype induces a complex association signal over a number of SNPs.

2.4. *Simulation approach*

For simulation studies of each gene, we generated cohorts of 100 000 individuals with haplotypes assigned at random from $\mathbf{H}$ in proportions f assuming Hardy–Weinberg equilibrium.

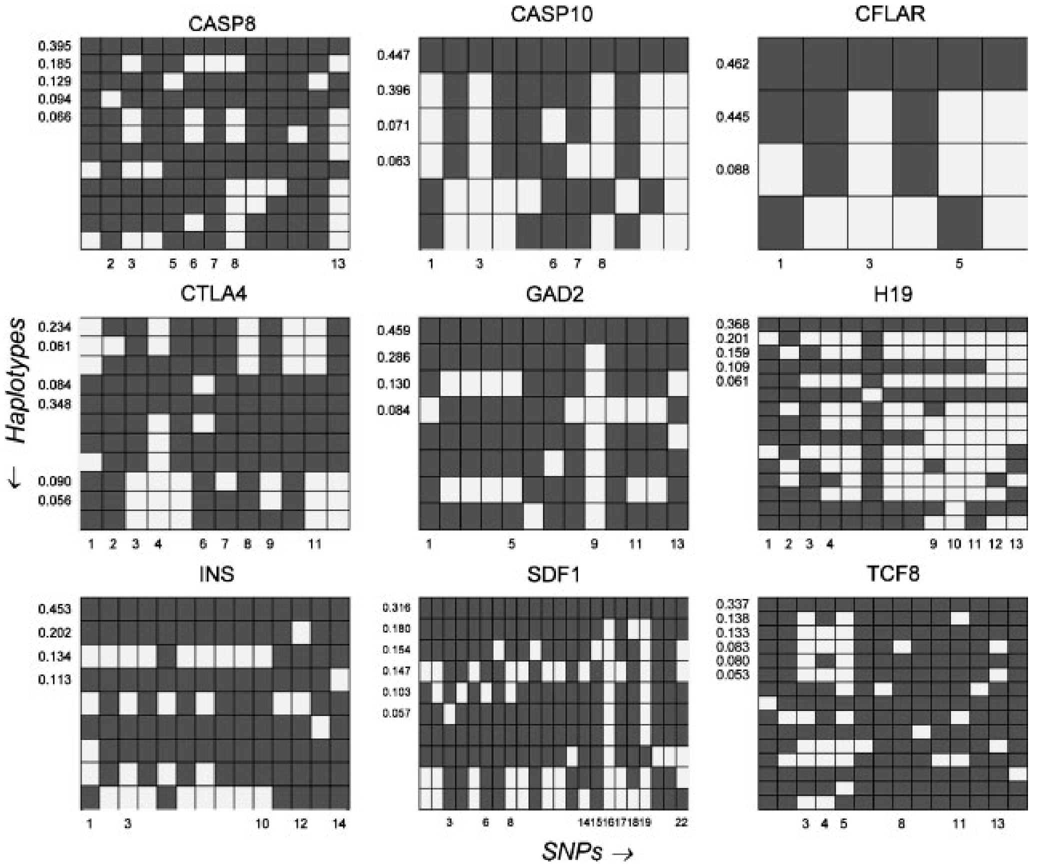

Figure 1. Haplotype structures in the nine-gene panel. Panels correspond to genes, rows to haplotypes, and columns to SNPs. Grey indicates that the SNP allele matches the corresponding allele in the referent haplotype in row 1. White indicates that the SNP allele differs from the corresponding allele in the referent haplotype. The y-axis labels show normalized haplotype frequencies $\geqslant 0.05$. Unlabelled rare haplotypes are pooled in a joint analysis of haplotypes. For each gene, the SNP allele frequencies π are determined from the binary haplotype matrix $\mathbf{H}$ and the haplotype frequency vector f by the formula $\pi = \mathbf{H}'f$. The x-axis labels identify the position of all SNPs with $\min(\pi_j, 1-\pi_j) \geqslant 0.05$. Labelled SNPs are included in SNP-based association analyses.

Then, for each individual, we generated a Bernoulli trial in which the probability of disease was specified according to the assumed linear logistic model. The baseline disease risk β_0 (a nuisance parameter) was set to 0.01 when $n_1 = n_0 = 300$.

Finally, we sampled n_1 cases and n_0 controls from each cohort to obtain a single case–control sample. The sampling was repeated a large number of times B so that the statistical power of the multiple test procedures could be estimated. The type I error rates were estimated by simulating studies under the complete null. The output data sets included case–control status, SNP genotypes, and true SNP haplotypes. Multiple test procedures were run on all SNPs and haplotypes in each gene with a minor allele frequency $\geqslant 0.05$, as indicated in Table I.

We used a similar approach to simulate data from all nine genes considered as a panel of independent candidates.

3. RESULTS

3.1. FWER control and power for tests of a single candidate gene

Table II shows type I error rates under the complete null for multiple test procedures applied to simulated case–control data from the nine-gene panel. The nominal type I error rate was 0.05. The false positive rate is very high for SNP-based analysis if no corrections are made for multiple comparisons, ranging from 18 to 45 per cent across the panel. In contrast, each multiple test procedure controls the type I error rate close to or below the nominal level.

Next, for each gene, we simulated a sequence of non-null situations where each SNP or haplotype in turn is associated with disease with $\mathrm{RR} = 1.5$. Figure 2 presents power curves when disease susceptibility is conferred by a SNP with a co-dominant effect, and Figure 3 presents power curves when disease susceptibility is conferred by a haplotype with a co-dominant effect. In these figures, we consider joint and marginal tests of SNPs and haplotypes.

When disease susceptibility was conferred by a SNP (Figure 2), the marginal test of SNPs had higher power than the marginal test of haplotypes, or else the power was similar. These results present the best-case scenario for haplotype analyses because here the linkage phase is known. There were differences in performance over and above the differences in the number of degrees of freedom. For example, in CFLAR and TCF8, both the number of tested SNPs and the number of tested haplotypes were equal, but the marginal test of SNPs dominated.

Table II. Single-gene analysis, type I error rates under the complete null.

	Multiple test procedure[*]					
Gene	SNP tests uncorrected for multiple comparisons[†]	Joint test of SNPs, $p^{\mathrm{SNP}\ddagger}_{\mathrm{Joint}}$	Marginal test of SNPs, $p^{\mathrm{SNP}\ddagger}_{\mathrm{Marg}}$	Joint test of haplotypes, $p^{\mathrm{Hap}\ddagger}_{\mathrm{Joint}}$	Marginal test of haplotypes $p^{\mathrm{Hap}\ddagger}_{\mathrm{Marg}}$	Omnibus test[§]
CASP8	0.331	0.057	0.040	0.055	0.051	0.054
CASP10	0.216	0.056	0.045	0.056	0.047	0.050
CFLAR2	0.177	0.061	0.044	0.061	0.041	0.056
CTLA4	0.258	0.055	0.041	0.055	0.048	0.048
GAD2	0.288	0.054	0.044	0.057	0.045	0.050
H19	0.230	0.060	0.038	0.055	0.049	0.054
INS	0.259	0.055	0.044	0.053	0.048	0.054
SDF1	0.336	0.056	0.043	0.054	0.047	0.046
TCF8	0.453	0.050	0.042	0.052	0.052	0.049

[*] Details of the procedures are given in Section 2. The nominal α level for each procedure was 0.05.

[†] In the uncorrected analysis, a gene is declared to be associated with disease if any single-SNP likelihood ratio trend test is significant at the nominal 0.05 level without adjustment for multiple comparisons. These ‘per-comparison error rates’ were estimated from $B = 10\,000$ replications.

[‡] These type I error rates were estimated from $B = 10\,000$ replications.

[§] For each replication, p_{omni} was estimated from 1000 permutations, and the type I error rate was estimated from $B = 1000$ replications.

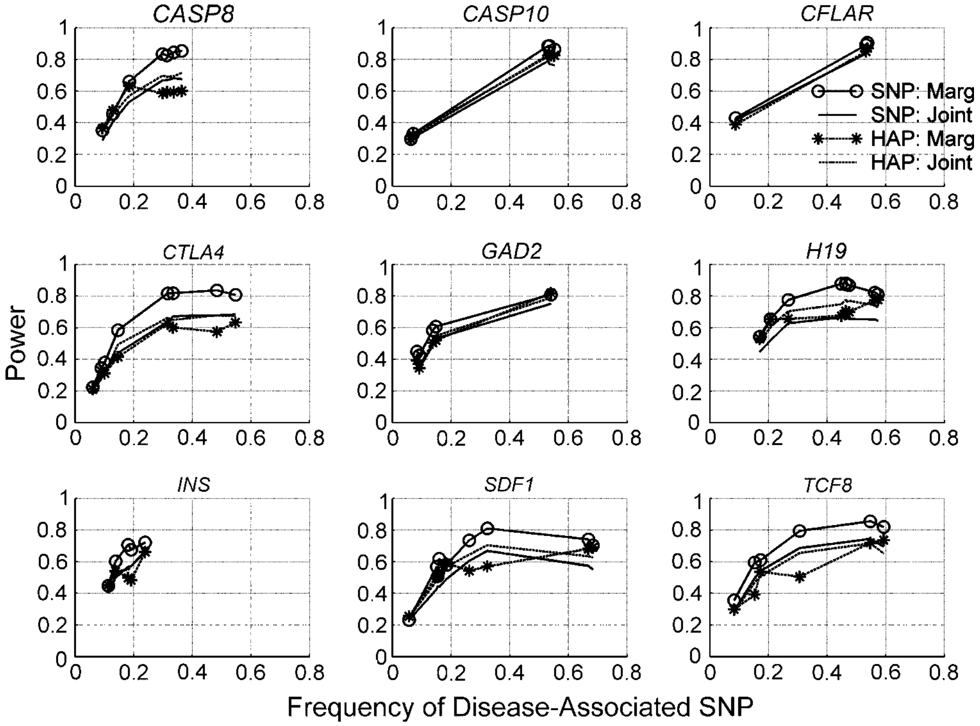

Figure 2. Power curves for marginal and joint genetic association tests when disease susceptibility is conferred by a SNP with a co-dominant effect. Panels correspond to genes, and ordinates to power. Within panels, each abscissa value indexes a situation, arranged from the least common to the most common SNP in the gene. In each situation, $RR = 1.5$ per copy of the variant SNP, and studies of $n_1 = n_0 = 300$ cases and controls were assessed. Power curves are shown for the marginal and joint tests of SNPs, and for the ideal marginal and joint tests of haplotypes based on phase-known data. For each situation, power was estimated from $B = 1000$ replications.

In all nine genes, the marginal test of SNPs was more powerful than the joint test of SNPs, but frequently, the marginal test of haplotypes was less powerful than the joint test of haplotypes.

Conversely, when disease susceptibility was conferred by a haplotype (Figure 3), the marginal test of haplotypes had higher power than the marginal test of SNPs, or else the power was similar. These scenarios were favourable to haplotype analysis, because, with the exception of CFLAR and TCF8, the genes had fewer haplotypes than SNPs, and the linkage phase was known. Nonetheless, in some configurations, the increase in power of the marginal test of haplotypes *versus* the marginal test of SNPs was substantial, as much as 38 percentage points. Therefore, in 14 situations where the power of the marginal test of haplotypes exceeded the power of the marginal test of SNPs by 5 or more percentage points, we also evaluated the corresponding marginal test of haplotypes applied to unphased SNP genotype data. In these 14 situations, the drop in power due to phase ambiguity was negligible, averaging 0.4 per cent. Therefore, after allowing for phase ambiguity in these situations, it remained the case that the marginal analysis of haplotypes dominated the marginal analysis of SNPs.

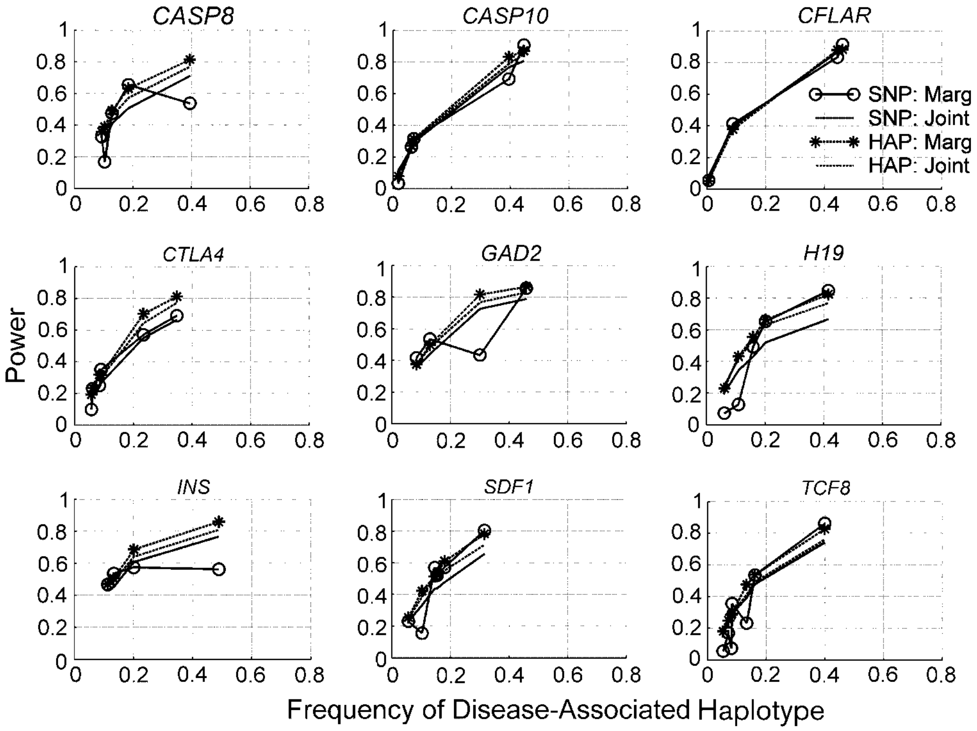

Figure 3. Power curves for marginal and joint genetic association tests when disease susceptibility is conferred by a haplotype with a co-dominant effect. See the legend to Figure 2 for details. Within panels, abscissa values are arranged from the least common to the most common haplotype in the gene.

Interestingly, in both Figures 2 and 3, the power curves for the joint analysis of SNPs and the joint analysis of haplotypes had similar shapes. Joint analysis of SNPs in frequently had higher power when disease susceptibility was conferred by a SNP (Figure 2), while joint analysis of haplotypes frequently had higher power when disease susceptibility was conferred by a haplotype (Figure 3). Presumably, in Figure 3, these curves would be more similar in magnitude if the linkage phase was unknown.

Next, we considered the power of the omnibus test combining the marginal test of SNPs with the joint test of haplotypes. When disease susceptibility was conferred by a SNP (Figure 4), the omnibus test closely tracked the marginal test of SNPs, which was the more powerful of the two component tests in these situations. Similarly, when disease susceptibility was conferred by a haplotype (Figure 5), the omnibus test closely tracked the joint test of haplotypes, which was generally the more powerful of the two component test in these alternative situations.

The trends apparent in Figures 2 and 3 were also seen in situations where disease susceptibility was conferred by a SNP or haplotype with a dominant effect (data not shown). For recessive models, the same qualitative findings were also observed, but as expected, the actual power to detect association is considerably lower.

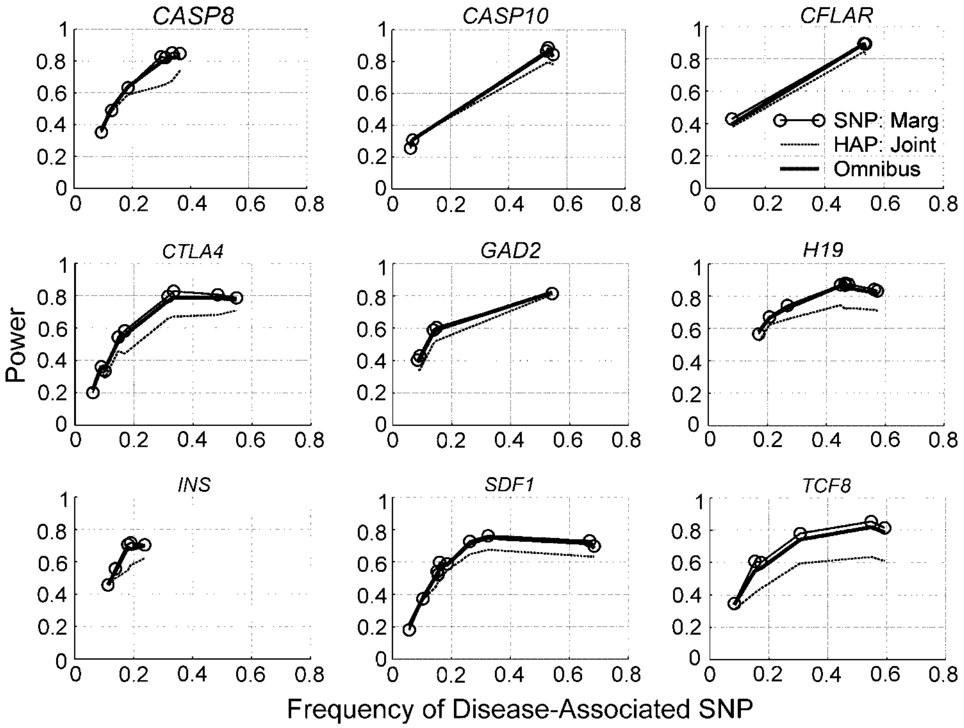

Figure 4. Power curves for the omnibus test when disease susceptibility is conferred by a SNP with a co-dominant effect. See the legend to Figure 2 for details. Power curves are shown for the marginal test of SNPs, the ideal joint test of haplotypes, and the omnibus test that combines these two. For each situation, power was estimated from $B = 1000$ replications of studies with $n_1 = n_0 = 300$ cases and controls; omnibus test p-values were estimated from 1000 permutations.

3.2. FDR control for tests of multiple candidate genes

We considered specific situations defined using the entire panel of genes, and assessed the actual FDR and power of the two-stage testing approach (Table III).

Each of the two-stage testing approaches controlled the expected proportion of falsely rejected genes. In theory, the expected gene-level FDR equals $(G_0/G)q$, where G_0 is the number of genes in the panel that are not associated with disease. The power of the approximate omnibus test was intermediate between the two-stage Simes and the two-stage haplotype analysis.

4. DISCUSSION

Case–control association analysis of multiple SNPs or haplotypes in a single gene presents a statistical multiple comparisons problem. So too does analysis of multiple genes in a panel of candidates. Statistical multiplicity problems will likely increase over time, as genotyping technologies advance and more genes and more variants within genes are probed. Ultimately, all

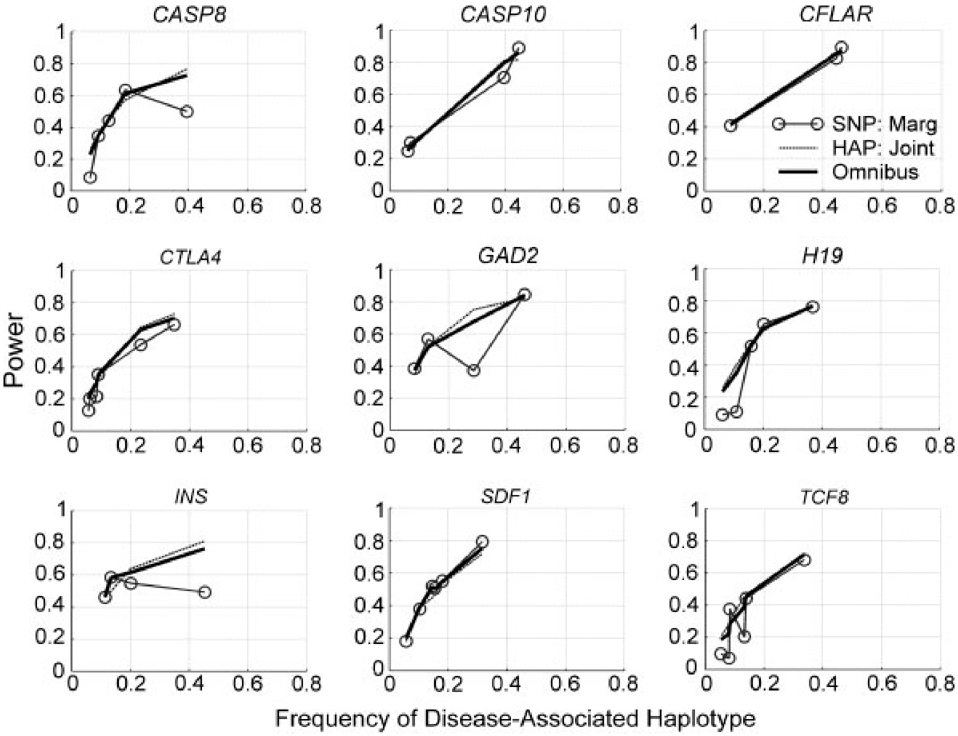

Figure 5. Power curves for the omnibus test when disease susceptibility is conferred by a haplotype with a co-dominant effect. See the legend to Figure 4 for details. Within panels, abscissa values are arranged from the least common to the most common haplotype in the gene.

the common variants within a gene or gene region could be investigated. This comprehensive strategy has been advocated as the approach of choice once the technology permits it [35]. On a small scale, this is precisely the strategy we investigated here. Assuming that our gene panel is representative, our results may provide some guidance about how to analyse such gene-based studies.

In our panel, when disease susceptibility was conferred by a SNP, the marginal test of SNPs had similar or higher power than the joint test of SNPs, the marginal test of haplotypes, and joint tests of haplotypes. Conversely, when disease susceptibility was conferred by a haplotype, the marginal test of haplotypes had similar or higher power than the other tests considered. Future studies are needed to elucidate circumstances under which the impact of phase ambiguity is large enough to negate the possible advantages of an ideal, phase-known haplotype analysis. However, in the limited number of situations considered here, there was only a minor average loss of power due to phase ambiguity. For situations where the genetic susceptibility mechanism is uncertain—and this may be most of the time—the omnibus test combining SNP and haplotype analysis appears to closely track the more powerful of the two component tests, which is generally unknown. In the light of our results, it appears that an omnibus test combining the marginal model of SNPs with the marginal model of haplotypes

Table III. FDR and power for two-stage gene-panel analysis.

		Gene-specific tests					
Situation[*]		Marginal test of SNPs		Joint test of haplotypes		Omnibus test[‡]	
		Power[†]	FDR[†]	Power	FDR	Power	FDR
0:	Complete null	0	0.044	0	0.052	0	0.038
1:	SNP5 : *CASP8* SNP3 : *CTLA4* SNP8 : *SDF1*	0.392	0.031	0.278	0.035	0.333	0.027
2:	SNP1 : *CASP10* SNP3 : *CFLAR2* SNP9 : *GAD2*	0.741	0.027	0.639	0.034	0.676	0.026
3:	SNP1 : *CASP10*	0.665	0.040	0.534	0.047	0.607	0.036
4:	HAP1 : *CASP8*	0.209	0.039	0.492	0.047	0.419	0.036
5:	HAP1 : *H19* HAP2 : *INS* HAP1 : *TCF8*	0.476	0.029	0.496	0.035	0.484	0.025
6:	HAP1 : *CASP8*↑ HAP2 : *CASP8*↓	0.845	0.037	0.917	0.047	0.901	0.035
7:	HAP2 : *GAD2* SNP4 : *TCF8*	0.628	0.031	0.471	0.041	0.568	0.029
8:	HAP2 : *GAD2*↑ SNP4 : *TCF8*↓	0.582	0.033	0.441	0.044	0.518	0.032

[*] The specified alleles increased disease risk by 1.5-fold per copy; as noted, in situations 6 and 8, haplotype 2 in *CASP8* and SNP 4 in *TCF8 decreased* the risk by 1.5-fold per copy.

[†] For each situation, Power and FDR were estimated from $B = 10\,000$ replications of studies with $n_1 = n_0 = 300$ cases and controls. Power was defined here as the expected proportion of truly associated genes that are found to be significant.

[‡] The omnibus test combined a marginal test of SNPs with a joint test of haplotypes using the Bonferroni correction.

would have been slightly more powerful in the tested situations where a single haplotype conferred susceptibility. We suspect the advantage of this particular omnibus test would be diminished if the susceptibility model was more complex.

Our results have implications for a haplotype-first strategy. Haplotype analysis can be very efficient in the laboratory when tag SNPs are selected [34, 36], allowing large genomic regions to be probed. However, as our examples illustrate, if haplotype analysis is negative for association, this does not necessarily mean that the gene is completely negative. Testing haplotypes first appears to be a sound strategy for a first-pass study, especially of large genomic regions, but it is clear that negative regions might require additional testing of SNPs. Hence, the comprehensive strategy we study here may ultimately be necessary.

Three key limitations of our simulation study must be noted. First, the nine-gene panel cannot be entirely representative of the human genome; Johnson and colleagues selectively resequenced the panel's genes over non-contiguous regions. Therefore, our haplotype structures might provide an idealized and perhaps overly simplistic model of LD. Second, the non-null situations we considered in Figures 2–5 might be overly simplistic, since only a single SNP or haplotype conferred susceptibility. Third, there is no guarantee that the comparative performance of the tests will be the same as indicated here, if only a limited number of SNPs are genotyped in a candidate gene. However, although the existence of non-typed SNPs might result in a loss of power, the tests we describe here should nonetheless control the type I error over the markers that are available for analysis.

For gene-panel analysis, we considered only the BH-FDR procedure for this report. Our main objective was to confirm, in a series of examples, the common-sense notion that the better the base test, the better the gene-panel analysis. One limitation of the BH-FDR procedure has been noted [37]. When the dependency assumptions are met, it is guaranteed to control the FDR on average over repeated experiments. However, it is not guaranteed to control the FDR within a set of rejected hypotheses from a particular experiment, called the conditional FDR. Procedures have been developed to control the conditional FDR in the context of large families of random independent test statistics with identical distributions for both the null and the non-null hypotheses [38]. In this setting, when the fraction of non-null hypotheses is substantial, conditional FDR procedures can be more powerful than unconditional FDR procedures. This approach has proven to be popular in microarray data analysis [39]. It is not yet clear whether these approaches are applicable to association analysis of moderate-sized gene panels with heterogeneous test statistics and a low prior probability of association. If the signal-to-noise ratio in a gene panel was thought to be high *a priori*, both conditional FDR and other unconditional FDR procedures might be more powerful than the BH-FDR procedure [40–42].

We see many opportunities to refine and extend both FWER strategies for candidate gene analysis, and FDR strategies for gene-panel analysis. For candidate gene analysis, it might be advantageous to augment the trend test to reflect model uncertainty [43]. We are developing a resampling-based multiple test procedure for SNPs that appears to provide 3 per cent higher power on average than the marginal test of SNPs studied here. The resampling-based test can also be combined with a haplotype analysis of unphased SNP genotype data to yield a sharper omnibus test. For gene-panel analysis, it would be helpful to have power and sample size guidelines. For a given FDR procedure and sample size, if the panel contains m_1 truly associated genes with given effect sizes, how many unassociated genes m_0 can also be considered before the expected proportion of associated genes that is detected falls to unacceptable levels?

For investigators with genetic association data in hand, multiple comparisons procedures can help to balance the desire for power against the implicit costs of false positives, and possibly help to diminish the prevalence of false positive reports in the literature [20, 44–46]. Perhaps the best solution to the multiplicity problem is to assemble good panels of candidate genes. For this purpose, substantive knowledge is essential, but statistical methods such as meta-analysis and Bayesian inference might also play a role.

ACKNOWLEDGEMENTS

We are indebted to the editor and to two anonymous reviewers for insightful and constructive comments. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

REFERENCES

1. Kruglyak L, Nickerson DA. Variation is the spice of life. *Nature Genetics* 2001; **27**(3):234–236.
2. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 2001; **29**(1):308–311.
3. The International HapMap Project. *Nature* 2003; **426**(6968):789–796.

4. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nature Genetics* 2001; **29**(2):229–232.
5. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**(5529):489–493.
6. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science* 2002; **296**(5576):2225–2229.
7. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends in Genetics* 2003; **19**(3):135–140.
8. Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *American Journal of Human Genetics* 2003; **73**(2):285–300.
9. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**(5547):1719–1723.
10. Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405**(6788):847–856.
11. Kwok PY. Genomics. Genetic association by whole-genome analysis? *Science* 2001; **294**(5547):1669–1670.
12. Clark AG. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Current Opinion in Genetics and Development* 2003; **13**(3):296–302.
13. Schork NJ. Power calculations for genetic association studies using estimated probability distributions. *American Journal of Human Genetics* 2002; **70**(6):1480–1489.
14. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Research* 2001; **11**(1):143–151.
15. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* 2002; **70**(2):425–434.
16. Stram DO, Leigh PC, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case–control study of unrelated individuals. *Human Heredity* 2003; **55**(4):179–190.
17. Epstein MP, Satten GA. Inference on haplotype effects in case–control studies using unphased genotype data. *American Journal of Human Genetics* 2003; **73**(6):1316–1329.
18. Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case–control studies. *American Journal of Human Genetics* 2003; **72**(5):1231–1250.
19. Schork NJ, Fallin D, Lanchbury JS. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics* 2000; **58**(4):250–264.
20. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends in Genetics* 2001; **17**(7):407–413.
21. McCarthy JJ, Hilfiker R. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nature Biotechnology* 2000; **18**(5):505–508.
22. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Statistics in Medicine* 1990; **9**(7):811–818.
23. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3): 751–754.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*) 1995; **57**(1):289–300.
25. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. *Proceedings of the National Academy of Sciences of the U.S.A.* 1999; **96**(26):15173–15177.
26. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity* 2003; **55**(1):56–65.
27. Samuel-Cahn E. Is the Simes improved Bonferroni procedure conservative? *Biometrika* 1996; **83**(4):928–933.
28. Krummenauer F, Hommel G. The size of Simes' global test for discrete test statistics. *Journal of Statistical Planning and Inference* 1999; **82**(1–2):151–162.
29. Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**(440):1601–1608.
30. Westfall PH, Zaykin DV, Young SS. Multiple tests for genetic effects in association studies. *Methods in Molecular Biology* 2002; **184**:143–168.

31. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 1999; **82**(1–2):171–196.
32. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**(4):1165–1188.
33. Sarkar SK. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics* 2002; **30**(1):239–257.
34. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identification of common disease genes. *Nature Genetics* 2001; **29**(2):233–237.
35. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics* 2004; **75**(3):353–362.
36. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human Heredity* 2003; **55**(1):27–36.
37. Zaykin DV, Young SS, Westfall PH. Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller *et al. Genetics* 2000; **154**(4):1917–1918.
38. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*) 2002; **64**(3):479–498.
39. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 2002; **23**(1):70–86.
40. Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 1999; **82**(1–2):163–170.
41. Kwong KS, Wong EH. A more powerful step-up procedure for controlling the false discovery rate under independence. *Statistics and Probability Letters* 2002; **56**(2):217–225.
42. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 2000; **25**(1):60–83.
43. Zheng G. Use of max and min scores for trend tests for association when the genetic model is unknown. *Statistics in Medicine* 2003; **22**(16):2657–2666.
44. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nature Genetics* 2001; **29**(3):306–309.
45. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large *versus* small studies: an empirical assessment. *Lancet* 2003; **361**(9357):567–571.
46. Tabor HK, Risch NJ, Myers RM. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics* 2002; **3**(5):391–397.